The use of TTS tapes in studies of computer typesetting practice

J. L. Dolby & R. C. Saunders

The use of computers in the printing industry raises the need for careful measurement of composition practice at a level that is seldom thought of when human intelligence plays an important part in the proceedings. The fact that a number of printers have been using TTS tape (which is, of course, a machine readable medium) for driving linecasters raises the possibility that some of the detail required can be accumulated with great speed and accuracy from the TTS tape directly without the need for time-consuming measurements.

Let us here consider some of the conversion problems necessary for a typical computing centre to make use of TTS tape and provide some illustrations of the information being gathered by the University of Newcastle upon Tyne computer typesetting project.

Six-hole advanced-sprocket TTS tapes were converted to eight-channel for input to the KDF9 computer. The conversion simply entailed a straight copy of the TTS tapes, but with channels 5 and 4 duplicated in channels 7 and 8. Therefore there can be no even or odd parity checking on these 8-hole tapes, but a programming check is used to ensure that channels 4 and 5 are always identical with 8 and 7, as read by the tape reader. All input routines used double-buffering with 512-word buffer sizes. To speed up the reading time, all information from the paper tapes was transferred to magnetic tape. The general outline of the input routines remained the same.

The internal coding is held in Q-store constants which can be altered to suit the program being developed. The KDF9 Q-store constants can hold three separate integers to store the information for the internal coding. Normally integers 1 to 26 are used for the alphabet, 27 for apostrophe, 28 for hyphen, 29 to 32 for ligatures, negative constants for 'terminators', eg space, carriage return or punctuation, and zero for codes to be ignored, eg thin space, if the particular program wishes to ignore these (as they are normally used between small caps at the beginning of a chapter). In the general case, with use of lower and upper rails on capital unshift and shift, the number of different codes is 63, so that 21 Q-store constants suffice to hold the information about the internal code for one of the four possible combinations. Hence 84 Q-store constants are required for the 252 possibilities. A marker is used to show the present state of shift and rail so that the relevant Q-store constant can be picked out.

Character frequency studies

The simplest kinds of information to be gathered involve the frequency of occurrence of the various characters, character pairs, character triplets, etc. Depending upon the use that is to be made of the data, these counts may be accumulated on a character code basis (ie a lower case **a** is counted separately from an upper case **A** or an italic lower case **a** etc.) or on what we might call a literary basis where an **a** is an **A** or an **a**. For studies of keyboarding practice and matrix case arrangement, the former case is more useful and, incidentally, easier to obtain.

In Fig 1 we show a brief count of the character codes for the first reel of paper tape used in printing *Goldfinger* (kindly lent to us for the purpose by Richard Clay & Sons). The results of the count are rather as one would expect: a total of 130 character codes actually occurred on this tape; the lower rail clearly dominates the upper rail by a considerable margin; and the letter frequencies follow the usual pattern. It is interesting to note that despite the large number of codes that were **not** used at all, it was still necessary to use the stop code thirteen times through the tape. This suggests that where the tape had been previously processed through a computer so that the entire set of characters required was known, there might be some economic gain in modifying the matrix case arrangement in order to eliminate the need for stop codes. Whether such a procedure would be viable in practice is another question.

Another observation that can be made is that the ligatures **ff, fi,** and **ffi** occurred 144 times out of the total count of 44,638 characters. (This information may well be affected by the fact that the chief character of the book contains the ligature **fi.**)

Of somewhat greater interest is the count of all possible character pairs. In Fig 2 we show a portion of this count where the character in the left hand column is the first character of the pair and the character across the top of the figure is the second character of the pair. Here we have amalgamated the upper and lower case and upper and lower rail counts into a single count, primarily to reduce the collection of figures to a reasonable number. Even with these reductions the entire set numbers $62 \times 62 = 3{,}844$ characters. Approximately half of these character pairs do not occur in the small sample these counts were made from, and a great many more pairs occur with very small frequencies.

Several uses for information of this kind are evident. Those studying keyboard arrangements, be they wholesale reorderings, as proposed by Dvorak, or minor changes in the non-literal keys and/or introduction of new keys for special functions such as shown in the Newcastle upon Tyne keyboard, the Purdy-

Fig 1 Frequency with which individual characters are used in Goldfinger*—Ian Fleming —as shown in TTS linecaster codes*

		LWR RAIL		UPR	
		UNSH	SH	UNSH	SH
1	T	2769	182	0	0
2	C RET	131	0	0	0
3	O	2631	41	8	0
4	SPACE	7603	5	5	4
5	H	2134	114	0	0
6	N	2087	24	6	0
7	M	634	48	2	0
8	ELEVATE	26	0	0	0
9	L	1422	7	1	0
10	R	1739	18	7	0
11	G	745	80	0	1
12	I	1821	122	8	0
13	P	523	18	3	0
14	C	752	25	0	0
15	V	250	1	5	0
16	THIN SP	17	44	18	35
17	FF 5	23	1	0	0
18	APOSTR'	222	91	1	0
19	FI 9	117	0	0	3
20	+THINSP	0	0	0	0
21	EM LDER	0	0	0	0
22	COMMA	314	1	0	0
23	FULL STOP	653	7	3	0
24	STOP CODE	10	0	1	2
25		1	0	0	3
26	* 4	0	0	1	4
27	; :	1	2	0	2
28	8	0	3	0	0
29	FL 0	20	6	7	1
30	ENSPACE	0	7	46	0
31	CENTER	5	0	0	0
32	E	3911	31	8	0
33	Z	34	0	0	0
34	D	1676	12	1	0
35	B	495	172	0	0
36	S	1849	77	0	2
37	Y	572	19	0	0
38	F	524	9	0	0
39	X	30	0	4	0
40	A	2327	65	10	0
41	W	811	48	0	0
42	J	38	10	0	1
43	SHIFT	1113	0	0	0
44	U	922	10	5	0
45	Q	34	2	3	0
46	K	309	12	0	0
47	UNSHIFT	1113	0	0	0
48	'QU' 3	52	0	3	7
49	) (	0	0	3	0
50	HYP	43	0	0	0
51	UPR ROW	64	0	0	0
52	EMSPACE	273	0	0	0
53	6	0	0	0	2
54	FLUSHLEFT	126	0	0	0
55	FFI 1	4	0	0	11
56	'EXCL'	0	17	1	0
57	DASH 2	2	1	0	6
58	BELL	0	0	0	0
59	LWR ROW	80	0	0	0
60	£ 7	0	1	0	3
61	EN LDER	0	0	2	4
62		0	0	0	0

McIntosh keyboard and the like, will presumably find these statistics of some use in determining what orderings and/or extra keys are likely to speed up the keyboarding job. Similarly, those interested in character recognition systems as input to computer operations should find the sequences of successive characters of some interest, particularly when it is necessary to recognize the short letters, such as **i** in their immediate context. A character recognition machine that has to recognize type in context may have difficulty in determining character boundaries. If this is the case, and the character recognizing device has a 'window' of fixed size, it may be necessary to consider all pairs that can fit in the window as if they were distinct characters. The pair **in,** for instance, will occupy less horizontal room than the single character **w** in most type fonts (and may even occur more frequently, though this was not the case in this sample where **w** occurred 1,725 times and the pair **in** occurred only 1,143 times).

Hyphenation and justification practice

There has been much discussion about the 'proper' rules for hyphenation and justification, stemming mainly from the basic difficulties of finding highly accurate rules for English writing practice. Although several companies have prepared algorithms of sufficient accuracy for some purposes (when a large backup store for exception words is provided) there is still considerable doubt as to what constitutes reasonable practice; how much this practice differs from practice in the United States; and what modifications can be made without justifiable complaint from the ultimate customer.

Clearly, no simple set of rules is likely to do the job. Hence some trade-off between hyphenation and better justification procedures seems called for. Some information on the hyphenation practice of the printing industry in the UK is being obtained. However, the measurement of interword spaces is time-consuming when it must be done by hand, and the accumulation of justification measurements made on TTS tapes will speed up the process of determining to what extent a human operator leans on hyphenation when he could in fact end the line without it.

Cast-off calculations

One of the nice side benefits from the use of a computer in a typesetting installation is the greater flexibility it provides for formatting the design of the book. Generally, most printers wish the total corpus to run to a multiple of thirty-two pages, or, if possible, a multiple of sixty-four pages. A number of choices are available to the compositor to obtain this result: choice of font, type and set size, number of lines per page, measure, frequency of hyphenation (and hence density of words to the line), treatment of footnotes, references, illustrations, beginning and ending material and so forth. In general, it is not reasonable to expect in the immediate future that a computer program of sufficient sophistication will be available to the industry to make this decision. However, it is reasonable to obtain much more precise measurements of the amount of pages that a given amount of text will occupy, once the various variables have been determined.

Operationally, this means that one could read the entire text into the computer along with certain variations of format and obtain the number of pages of set material for each combination. As the calculations for accomplishing this are relatively straightforward, one could probably obtain as many as a dozen different analyses in a single pass of the text through the computer. If none of the given alternatives was satisfactory, one could then make a second pass with the knowledge gained from the first run in hand and in all likelihood have the required answer. The testing of such a routine would obviously require test runs on a number of books to ensure that the various approximations necessary to the cast-off calculations are of sufficient accuracy. At the same time, the availability of quantities of TTS tape will be of considerable use in establishing the form of the approximations in the first place.

Linguistic studies

It is difficult to discuss the utility of large quantities of written text in machine readable form without referring to other problems that, at least on the surface of the matter, are not directly related to computer typesetting. As computational linguistics moves on to more difficult problems, it is inevitable that linguists will place more and more demands on the printing industry to supply sample tapes for use in particular areas of research. Thus far, the primary area of linguistic development of interest to the printing industry has been the development of adequate hyphenation rules and in light of the time taken to accomplish this the printing industry may well ask how long before they may expect other useful studies from their point of view.

One obvious procedure to implement, when sufficient knowledge is available, is the problem of providing an index of terms used at the back of technical books and journal articles. It will probably be another five to ten years before procedures at that level of sophistication will be available. Meanwhile, large-scale tests of various hypotheses will be necessary and to the extent that material in machine readable form is available, this research will proceed at a faster pace.

Of more immediate use to the industry will be the development of automatic editing rules, which will again be greatly accelerated by the availability of large quantities of text in machine readable form. In any case, the industry will find an ever increasing number of demands for such material and although it may be too early to think in terms of a national depository for such material, individual printers should give the problem some thought so that appropriate contracts will be available to provide proper protection for the printer and publisher while allowing sufficient material to be made available so that workers in the field will be able to derive the necessary information for testing and development purposes.

To those interested in experimenting on the processing of the English language and/or the development and testing of various

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	-	143	232	345	2	28	80	3	295	1	67	402	168	1074	8	111	1	486	539	656
B	127	9	-	-	286	-	-	-	61	1	-	126	3	-	137	-	-	70	14	2
C	311	1	12	-	252	-	-	281	63	-	198	65	-	-	261	-	-	84	2	86
D	59	2	1	34	307	2	14	2	145	1	3	22	1	40	226	-	2	69	112	1
E	364	10	138	833	261	59	29	20	83	-	15	290	114	593	35	69	12	982	439	239
F	66	1	1	-	74	-	-	-	-	-	-	-	-	-	195	-	-	84	2	77
G	128	-	-	-	237	-	17	248	56	-	4	75	-	24	147	-	-	67	28	1
H	784	5	-	1	1982	2	-	-	617	-	-	6	4	4	252	2	-	78	4	175
I	63	11	295	216	100	73	233	2	4	-	45	207	178	1143	164	41	4	170	488	555
J	6	-	-	-	9	-	-	-	2	-	-	-	-	-	16	-	-	-	-	-
K	6	-	-	-	233	5	1	-	77	-	-	14	1	40	5	-	-	-	46	2
L	245	5	12	321	452	42	1	-	269	-	27	403	19	1	260	13	-	8	49	34
M	181	40	1	-	441	4	-	-	177	-	-	6	32	8	138	61	-	-	32	-
N	109	6	107	1081	411	8	690	2	126	1	84	34	4	28	196	2	4	2	132	483
O	35	42	62	90	22	441	17	6	75	3	120	242	255	1097	276	109	-	522	125	226
P	140	1	-	-	191	1	-	37	81	-	1	133	2	-	110	79	-	107	35	22
Q	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R	246	11	26	171	766	10	26	15	275	-	47	58	40	70	392	26	-	74	190	148
S	182	-	57	-	350	4	4	213	221	1	30	66	45	26	150	56	5	-	182	473
T	243	2	50	2	473	3	-	1842	339	-	-	73	9	-	619	-	-	143	70	115
U	38	34	45	16	56	1	106	-	78	-	1	217	36	200	3	120	-	236	220	348
V	24	-	-	-	395	-	-	-	58	-	-	1	-	-	32	-	-	-	-	-
W	480	1	1	8	202	1	-	166	221	-	-	11	-	104	185	-	-	14	20	-
X	9	-	7	-	7	-	-	-	34	-	-	-	-	-	-	32	-	-	-	25
Y	13	3	5	1	76	-	-	-	26	-	-	2	7	3	178	5	-	2	45	16
Z	3	-	-	-	34	-	-	-	5	-	-	-	-	-	2	-	-	-	-	1
FF	2	-	-	-	32	-	-	-	-	-	-	-	-	-	2	-	-	1	2	-
FI	-	-	2	2	8	8	4	-	-	-	-	15	-	117	-	-	-	24	3	6
FL	16	-	-	-	13	-	-	-	8	-	-	-	-	-	18	-	-	-	-	-
FFI	-	-	16	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-

Fig 2. Digram frequency counts (part of) taken from Goldfinger*—Ian Fleming—as shown in TTS linecaster codes*

procedures for linguistic data processing—including computer typesetting—the existence of large quantities of English text in machine readable form can be a great boon. Various conversion facilities must be established to take into account both the physical configuration of the tape itself and the compatibility of the TTS codes with internal computer representations. Equally, various degrees of sophistication in computer programming both in terms of file handling routines and in computer languages are necessary in most computing centres to take advantage of TTS material.